Knowledge Discovery and Data Mining: Project Proposal

By:

1. Preet Jhanglani
   CWID: 10474322
   Section A

2. Girish Budhrani
   CWID: 10477624
   Section A

3. Anirudh Jeevan
   CWID: 10475896
   Section A

4. Surya Giri
   CWID: 10475010
   Section B

Abstract:

Ever since the introduction of the concept of Big Data, we have had access to an endless amount of data being collected and waiting to be processed. Today, with the help of developments in the field of Knowledge Discovery and Data Mining techniques, we can develop various methodologies to classify and/or predict outcomes of an event based on the factors that influence it. These factors are widely available in the form of unorganized and unprocessed data. We can use knowledge discovery techniques to mine data for information, which can then be used to to answer questions pertaining to the data set. One such technique is Classification. Classification is a supervised learning method, where known and labelled data is used in order to learn to classify outcomes of unknown data into different categories or "classes". Our goal of this project is to use such Classification methods or some equivalent of it on a dataset with binary classification to predict whether the outcome a chosen event (or column) will be a 1 or 0.

The dataset chosen for this project is Company Bankruptcy Prediction, from Kaggle. The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. The problem pertaining to this dataset is to find whether a company (represented by rows) is on the verge of Bankruptcy (outcome: 1) or not (outcome: 0). The dataset has 96 columns. However, not all columns are factors that contribute towards the prediction of the outcome and hence need to be excluded. We will execute some dimensionality reduction procedures in order to filter the dataset for useful information and exclude the rest.

Once the data has been properly cleaned, we will analyze the following algorithms on the basis of the accuracy of their classification/prediction of bankruptcy:

1. Logical Regression

   Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

2. k-Nearest Neighbors

   k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation.

3. Naive Bayes

   Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

4. Random Forest

   Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

5. Artificial Neural Networks

   Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems inspired by the biological neural networks that constitute animal brains.

Link to the dataset:

https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction